人工智能决策公平感知的影响因素及其机制*

蒋路远[1] 曹李梅[1] 秦昕[1] 谭玲[2] 陈晨[1] 彭小斐[1]

([1] 中山大学管理学院, 广州 510275) ([2] 广东工业大学管理学院, 广州 510520)

摘　要 不平等问题是全球社会和经济发展需要应对的首要挑战，也是实现全球可持续发展目标的核心障碍。人工智能(artificial intelligence, AI)为缓解不平等、促进社会公平提供了新的途径。然而，新近研究发现，即使客观上 AI 决策具有公平性和准确性，个体仍可能对 AI 决策的公平感知较低。因此，近年来越来越多的研究开始关注 AI 决策公平感知的影响因素。然而，目前研究较为分散，呈现出研究范式不统一、理论不清晰和机制未厘清等特征。这既不利于跨学科的研究对话，也不利于研究者和实践者对 AI 决策公平感知形成系统性理解。基于此，通过系统的梳理，现有研究可以划分为两类：(1)AI 单一决策的公平感知研究，主要聚焦于 AI 特征和个体特征如何影响个体对 AI 决策的公平感知；(2)AI-人类二元决策的公平感知研究，主要聚焦于对比个体对 AI 决策与人类决策公平感知的差异。在上述梳理基础上，未来研究可以进一步探索 AI 决策公平感知的情绪影响机制等方向。

关键词 人工智能，算法，公平，决策，影响因素

1 引言

不平等问题是全球社会和经济发展需要应对的首要挑战。联合国经济与社会事务部所发布的《2020 年世界社会报告(World Social Report, 2020)》指出，全球不平等现象正在日益加剧，全球超过三分之二人口所生活的国家都面临着不平等问题进一步扩大的现状。这不仅会影响经济发展和社会稳定，还会阻碍全球可持续发展目标的实现(郑功成, 2009; Vinuesa et al., 2020)。人工智能(artificial intelligence, AI)在缓解不平等问题、促进社会公平等方面被寄予厚望 (曹培杰, 2020; Dalenberg, 2018; Loehr, 2015)，因为 AI 能够依靠大数据和复杂计算作出客观判断(徐鹏, 徐向艺, 2020; Lindebaum et al., 2020)，有助于减少人类决策过程的主观偏差和不公平(房鑫, 刘欣, 2019; Miller & Keiser, 2021)。例如，在应对新冠病毒方面，AI 有效缓解了人类偏见可能导致的医疗资源分配不公等问题(Bigman et al., in press)。

收稿日期：2021-7-13

* 本文受国家自然科学基金面上、青年项目(71872190, 71702202, 71802203)、教育部长江学者奖励计划青年项目、中央高校基本科研业务费专项资金(19wkpy17)的资助。

通信作者：秦昕，E-mail: qinxin@sysu.edu.cn。

尽管 AI 决策在促进公平方面成效突出并蕴藏着巨大潜力，但部分新近研究发现，即使 AI 和人类的决策结果在客观上没有差异，甚至优于人类，个体仍可能表现出算法厌恶(algorithm aversion; Burton et al., 2020)，并宁愿以牺牲效率为代价来反对 AI 决策(谢小云 等, 2021; Bigman & Gray, 2018; Longoni et al., 2019)。这是因为对 AI 决策的一系列假设(例如，去情境化; decontextualization)让个体认为 AI 决策比人类决策更不公平(Helberger et al., 2020; Newman et al., 2020)。例如，在同样的精确度下，个体认为，相比于 AI 决策，由心理学家进行犯罪风险预测和决策更公平，甚至认为按照强制性的指导政策决策也比 AI 决策更公平(Wang, 2018)。

上述研究表明，个体对 AI 决策的公平感知是研究者亟需关注的重要问题(张志学 等, 2021)。在理论上，探索个体对 AI 这一新兴决策主体的公平感知扩展了传统公平研究的视角(Newman et al., 2020; Ötting & Maier, 2018)；在实践上，探索 AI 决策公平感知的影响因素能够帮助优化 AI 决策系统中公平准则的构建与设计(谢洪明 等, 2019; Barabas et al., 2020; Helberger et al., 2020)。基于此，越来越多研究开始关注 AI 决策公平感知的影响因素及其作用机制。

自 2017 年以来，AI 决策公平感知的相关研究呈现加速增长的趋势，并吸引了越来越多社会科学和计算机科学领域的研究者投入其中。然而，目前 AI 决策公平感知的相关研究散布于不同学科领域，呈现出研究范式不统一、理论不清晰和机制未厘清等特征。这既不利于跨学科对话，也不利于研究者和实践者对 AI 决策公平感知形成系统性理解。基于此，本文系统地梳理了 AI 决策公平感知的相关研究，总结了 AI 决策公平感知的影响因素和机制等，以期为未来研究提供参考和启发。具体地，本文全面搜索了 Web of Science，PsycINFO，IEEE Xplore，Scopus 和中国知网 5 个主要数据库后获得 53 篇 AI 决策公平感知相关论文并对其进行系统性整理。鉴于公平感知是个体对决策主体公平性的主观感受(李超平, 时勘, 2003; 郭秀艳 等, 2017; Colquitt & Zipay, 2015)，参考以往研究(Karam et al., 2019; Rupp & Cropanzano, 2002)，本文根据决策主体将 AI 决策公平感知的相关研究分成两类：第一，AI 单一决策的公平感知研究，即决策主体仅涉及 AI，这部分研究聚焦于探讨 AI 特征、个体特征等如何影响个体对 AI 决策公平性的知觉。第二，AI-人类二元决策的公平感知研究，即决策主体涉及 AI 和人类两方，这部分研究聚焦于对比个体对 AI 与人类决策公平感知的差异。基于以上分类，本文系统性地梳理了 AI 决策公平感知的相关研究及其理论机制，并总结出现有研究的整体框架，进而为未来研究提供有价值的指引和建议。

2 AI 单一决策的公平感知研究

AI 单一决策的公平感知研究主要探讨 AI 作为决策者时，AI 决策公平感知的影响因素及其作用机制，包括 AI 特征和个体特征两个方面。

2.1 AI 特征

Lee 等(2019)所提出的算法决策公平理论框架认为，在 AI 决策过程中，透明性(transparency)、可控性(controllability)和规则性(rule)是影响个体公平感知的重要因素。基于上述框架，并结合现有文献，本文从 AI 的透明性、可控性、规则性和适当性(appropriateness)四个方面对现有文献进行回顾(见表 1)。

2.1.1 透明性

透明性是指个体对 AI 决策过程和结果等相关信息的可获得程度(Shin, 2020; Shin & Park, 2019)，包括透明度(Lee et al., 2019)、解释性(explainability; Shin, 2021b; Shin, in press; van Berkel et al., 2019)和解释风格(explanation styles; Binns et al., 2018; Dodge et al., 2019)等。AI 决策的透明性主要通过个体的感知可理解性影响其公平感知。现有研究表明，AI 决策的透明性对个体的公平感知可能存在双刃剑效应：既可能提高交互公平(interactional justice; Wang et al., 2020)，也可能降低分配公平(distributive justice; Lee et al., 2019)。一方面，透明性可以提高个体对 AI 决策的交互公平感知(Wang et al., 2020)。AI 决策透明性越高，所提供的解释越相关、具体和充分，个体对 AI 决策的理解程度越高，进而交互公平感知越高(Binns et al., 2018; Langer et al., 2018; Shin, in press)。例如，Binns 等(2018)发现，在 AI 评估汽车保险的决策情境中，向参与者展示的决策条件越透明，参与者对 AI 决策的交互公平感知越高。另一方面，透明性也可能降低个体对 AI 决策的分配公平感知(Lee et al., 2019)。例如，Lee 等(2019)发现，在群体分配情境下，AI 分配决策的相关信息(例如，分配结果等)的透明性越高，向群体内个体暴露出的分配不均匀的程度也越高，从而降低个体的分配公平感知(Lee et al., 2019)。

2.1.2 可控性

可控性是指个体对 AI 决策过程和结果的控制程度(Lee et al., 2019)，包括决策参与(locus of decision; Uhde et al., 2020)、算法开发(algorithm development; Wang et al., 2020)和决策控制(decision control; Lee et al., 2019)等。AI 决策的可控性通过满足个体需求(例如，个体对输入数据或决策逻辑的控制程度)来影响其公平感知。AI 决策的可控性越高，越有利于个体形成对他们有利的决策结果，从而越可能提高其公平感知(Lee et al., 2019; Wang et al., 2020)。例如，Uhde 等(2020)发现，在进行医院排班决策时，相比于仅由 AI 来进行排班决策，医护人员参与其中、与 AI 一起协作决策，会带来更高的公平感知。

2.1.3 规则性

规则性是指 AI 决策的规则取向(Hutchinson & Mitchell, 2019; Shin, 2020)，包括分配规范

[1](allocation norms; Saxena et al., 2020)、公平取向[2](fairness approaches; Cheng et al., 2021)和收益权衡[3](benefit trade-offs; Srivastava et al., 2019)等。AI 决策的规则性主要通过规则取向的一致性(即，AI 规则取向和个体规则取向的一致程度)来影响个体的公平感知。公平相关研究认为，个体基于平等、公平和需求三类规则取向评估分配公平性(Deutsch, 1975; Leventhal, 1976)。因此，当 AI 决策的规则越符合上述公平规则，个体对 AI 决策的公平感知就越高(Chang et al., 2020; Cheng et al., 2021)。例如，在再犯风险评估(recidivism risk assessment)和医学预测情境中，AI 决策的规则设置为平等规则取向时(即人人平等规则，demographic parity)，个体认为该决策的歧视性越小，公平感知越高(Srivastava et al., 2019)。

2.1.4 适当性

适当性是指 AI 决策所使用信息的适宜程度(Kasinidou et al., 2021)，包括决策因素适当性(appropriateness of factors; Kasinidou et al., 2021)、种族和性别等个体信息使用(information about race or gender; Harrison et al., 2020；Nyarko et al., 2020; Plane et al., 2018)和信息相关性(relevance; Grgić-Hlača, Redmiles, et al., 2018)等。AI 决策的适当性主要通过个体道德直觉影响个体的公平感知。当 AI 决策所使用的信息适宜程度较低(例如，采用性别、种族等具有偏见的信息)时，可能会唤醒人类潜在的、直觉性的道德基础，并认为这一决策有违公平(Nyarko et al., 2020)，进而降低其公平感知。例如，Harrison 等(2020)在探究是否准予刑事被告人保释时发现，当 AI 决策不使用种族信息时，个体会觉得 AI 决策更公平。

2.2 个体特征

部分研究探讨了个体特征对 AI 决策公平感知的影响(Htun et al., 2021; Saha et al., 2020; van Berkel et al., 2021)。道德感知(例如，对公平和歧视的感知等)在不同个体之间存在差异(陈晨, 张昕等, 2020; 周浩, 龙立荣, 2007; Graham et al., 2013)，即个体的人口统计学特征(例如，性别、年龄和政治观点等)和人格与价值观会对个体的道德直觉产生影响，进而影响个体对 AI 决策的公平感知(Graham et al., 2013; Pierson, 2017)。基于此，本文从个体的人口统计学特征与人格和价值观两方面来回顾个体特征对 AI 决策公平感知的影响研究。

2.2.1 人口统计学特征

现有研究主要探讨了性别、年龄、受教育程度和技术知识等对 AI 决策公平感知的影响。在性

[1] 分配规范包括分配原则(即平等取向、需求取向和公平取向等)和分配比例(即最优决策、比例决策和平均决策等)。
[2] 公平取向是指三种 AI 公平取向：无意识(不含种族等敏感信息)、统计均等(敏感属性的组间正分类率[the positive classification rates between groups]相等)、平均比例(敏感属性在各组间的假阳性率和假阴性率[the false positive and false negative rates between groups]相等)。
[3] 收益权衡是指基于利益衡量的 AI 决策标准。

别方面，现有研究发现 AI 决策的公平感知存在性别差异(Grgić-Hlača et al., 2020; Pierson, 2017; Pierson, 2018)。例如，Pierson(2017)发现，女性比男性更注重伦理道德，因此，对于采用了具有偏见信息(例如，性别)的 AI 决策，女性比男性的公平感知更低。在年龄方面，大部分研究发现年龄对 AI 决策公平感知没有影响(Saha et al., 2020; van Berkel et al., 2021)。在受教育程度方面，现有研究发现了不一致的结论。部分研究发现，个体的受教育水平越高，对 AI 原理的理解程度越高，公平感知越高(Saha et al., 2020)。另外一部分研究则发现了相反的结果，例如，van Berkel 等(2021)发现，受教育水平越高的个体，越倾向于对复杂的 AI 决策产生质疑，进而公平感知越低。在技术知识方面，Schoeffer 等(2021)发现，个体关于 AI 的基础知识水平越高，越能更好地理解 AI 决策背后的逻辑，公平感知也越高。

2.2.2 人格和价值观

现有研究主要从人格特质、个体政治倾向、公平信念和个体相关经历等方面探讨了人格与价值观对 AI 决策公平感知的影响。在人格特质方面，现有研究主要关注了大五人格、自我效能感、隐私关心等对 AI 决策公平感知的影响(Araujo et al., 2020; Htun et al., 2021; Langer et al., 2018)。例如，Htun 等(2021)发现，个体开放性越高，越觉得基于部分团队成员偏好的 AI 决策是可接受的，因此公平感知越高。Araujo 等(2020)发现，个体的隐私关心越高，越可能质疑 AI 决策的道德性，进而公平感知越低。在政治倾向方面，Grgić-Hlača 和 Zafar 等(2018)发现，政治倾向和个体对 AI 决策的公平感知显著相关，自由主义倾向的个体公平敏感度更高，对决策中包含政治倾向等可能带有偏见特征信息的 AI 决策的公平感知更低。在公平信念方面，研究发现，公平信念更高的个体会更追求公平，更倾向于认为 AI 决策是客观的，公平感知更高(Araujo et al., 2020; Shin, 2021a)。在个体经历方面，研究发现 AI 决策公平感知会受到个体以往相关经历的影响(Smith, 2020)。例如，参加过保释听证会的个体，会觉得 AI 决策使用的被告人信息并不能完全预测犯罪风险，因此对 AI 决策的公平感知较低(Grgić-Hlača et al., 2020)。

2.3 小结

综上，AI 单一决策的公平感知研究分别从 AI 特征(即 AI 决策的透明性、可控性、规则性、适当性)和个体特征(即人口统计特征和人格价值观)两方面探索了其对个体 AI 决策公平感知的影响。上述因素主要通过三种认知机制发生作用，即可理解性、需求满足和道德直觉。可以看出，这一类研究将 AI 这一新兴决策主体纳入公平研究中，借鉴传统公平研究的相关理论(例如，Qin et al., 2015)，来解释决策者(即 AI)和决策接受者的不同特征如何影响个体对 AI 决策的认知，并进一步影响其公平感知。这类研究丰富了我们对个体如何形成 AI 决策公平感知这一问题的理解。

表 1 AI 单一决策的公平感知研究总结

类别	文献数量	亚类别	机制	作者和年份
AI 特征	8	透明性	可理解性/需求满足	Binns et al., 2018; Dodge et al., 2019†; Lee et al., 2019†4; Shin, 2020; Shin, 2021b†; Shin, in press†; Shin & Park, 2019†; van Berkel et al., 2019†
	3	可控性	需求满足	Lee et al., 2019†; Uhde et al., 2020†; Wang et al., 2020†
	4	规则性	需求满足	Chang et al., 2020†; Cheng et al., 2021†; Saxena et al., 2020†; Srivastava et al., 2019†
	5	适当性	道德直觉	Harrison et al., 2020†; Grgić-Hlača, Redmiles, et al., 2018†; Kasinidou et al., 2021†; Nyarko et al., 2020†; Plane, et al., 2018
个体特征	6	人口统计特征	道德直觉/可理解性	Grgić-Hlača et al., 2020†; Pierson, 2017†; Pierson, 2018†; Saha et al., 2020; Schoeffer et al., 2021; van Berkel et al., 2021†
	6	人格和价值观	道德直觉	Araujo et al., 2020†; Grgić-Hlača, Zafar, et al., 2018†; Htun et al., 2021†; Langer et al., 2018†; Shin, 2021a†; Smith, 2020†

注：†指该论文使用了相关逻辑阐述机制，但未进行实证机制测量和检验。

3 AI-人类二元决策的公平感知研究

AI-人类二元决策的公平感知研究主要探讨了 AI 与人类作为决策主体时，个体的公平感知差异。Gray 等(2007)提出的心智感知理论(theory of mind perception)认为，个体通过感知感受性(perceived experience; 即感知情绪、痛苦、快乐等的能力)和能动性(perceived agency; 即思考、计划、自主行动的能力)两方面形成对各类实体(例如，人、动物、AI 等)心智(mind)的主观感知(杨文琪 等, 2015)，这一感知又会进一步影响个体对它们的评估和反应(Schein & Gray, 2018)。基于此，现有研究主要从感受性和能动性两方面来对比个体对 AI 和人类的主观感知差异，进而探究其对 AI-人类二元决策的公平感知影响。具体地，感知感受性主要体现在个体对该实体(即 AI 和人类)在决策过程中所展现出的表达和感知情绪、展现友好等能力的感知。基于此，部分研究从感知感受性差异(即机械属性 vs. 社会属性)来对比 AI 和人类决策公平感知的差异。感知能动性主要体现在个体对该实体(即 AI 和人)在决策过程中所展现出的自主能力的知觉，而决策的准确性和一致性是判断其能力的重要因素(Fischhoff & Broomell, 2020)。基于此，部分研究分别从决策的准确性(即简化属性 vs. 复杂属性)和一致性(即客观属性 vs. 主观属性)两个方面来对比 AI 和人类决策公平感知的差异(见表 2)。

3.1 机械属性 vs. 社会属性

[4] Lee 等(2019)同时探讨了 AI 特征的透明性和可控性，故归入两个亚类别。

个体对 AI 决策的机械属性认知是影响其对 AI 与人类决策公平感知差异的重要作用机制。决策过程的社会性(例如，情感和互动)是个体感知公平性的重要方面(秦昕 等, 2019; 吴燕, 周晓林, 2012; Qin, Huang, et al., 2018)。相比于人类决策，个体认为 AI 决策是缺少情感和互动的，因此认为人类决策比 AI 决策更公平(Helberger et al., 2020; Noble et al., 2021)。

3.1.1 情感

相比于人类决策，个体往往认为 AI 决策过程是缺少情感的(Martínez-Miranda & Aldea, 2005)，相比于人类决策，AI 决策具有更低的人情味(personableness)和善意(benevolence)，进而公平感知更低(Kaibel et al., 2019; Langer et al., in press)。例如，Helberger 等(2020)发现，面对“你认为谁将做出更加公平的决定：人类还是 AI/计算机，请解释原因”这一开放性问题，25.4%的参与者明确提到情感是影响决策公平感知的重要因素，而 21.9%的参与者认为 AI 是缺少情感的，人类决策更公平。进一步地，个体的相关经历会影响个体对传统人类决策方式的情感体验，从而调节了不同决策主体对决策公平感知的影响。例如，对于没有经历过人类决策偏见和招聘歧视的应聘者来说，相比于人类决策，AI 决策是缺乏情感的，因此认为 AI 决策更不公平；然而，对于经历过人类决策偏见和招聘歧视的应聘者来说，AI 决策比人类决策具有更高的人情味和表现机会(opportunity to perform)(Kaibel et al., 2019)，进而认为 AI 决策更公平。类似地，Lee 和 Rich(2021)也发现，对于拥有良好的医疗体验的个体来说，医生决策比 AI 决策更公平；但是，当个体对医生决策方式缺乏信任时，他们对 AI 和医生决策的公平感知并没有显著差异。

3.1.2 互动

相比于人类决策，个体感知 AI 决策具有更低的互动性、人际接触和尊重 (Acikgoz et al., 2020; Ötting & Maier, 2018; Schlicker et al., in press)，进而导致更低的公平感知。相比于人类面试，个体对 AI 面试感知到更少的表现机会和复议的机会(reconsideration opportunity)、更差的待遇(treatment)和双向沟通(two-way communication)，以及更低的提问适当性(propriety of questions)，这些因素会显著影响应聘者对 AI 面试的公平感知(Noble et al., 2021)。类似地，Nørskov 等(2020)发现，相比于人类视频面试，应聘者对 AI 面试的程序公平(procedural justice)和交互公平感知更低。进一步，AI 和个体的特征影响了个体对 AI 决策过程互动性的感知，进而影响了 AI 与人类决策的公平感知差异(Lee & Baykal, 2017)。例如，在 AI 特征方面，Wang(2018)发现，在司法决策中，由于个体认为使用 AI 决策缺乏对决策接受者的尊重，所以认为 AI 决策没有心理学家决策公平；但是增加 AI 决策的准确度和透明度、减少结果偏见，可以缩小由于互动因素带来的 AI 和人类决策的公平感知的差异。在个体特征方面，个体在人际互动中的影响力(即人际权力)调节了决策主体对个体公平感知

的影响。对于高人际权力的个体来说，在群体互动中能够获得控制感，因此，他们认为人类决策更公平；但对于人际权力较低的个体来说，AI 和人类决策的公平感知没有显著差异(Lee & Baykal, 2017)。

3.2 简化属性 vs. 复杂属性

个体对 AI 决策的简化属性知觉也是影响其对 AI 和人类决策公平感知差异的重要内在机制(Höddinghaus et al., 2021; Lee, 2018; Suen et al., 2019)。现有研究发现，个体认为 AI 决策结果是根据有限的数据进行统计拟合得到的，因此，个体会认为 AI 决策忽视了背景和环境知识(Balasubramanian et al., in press)，简化了信息处理过程。Newman 等(2020)发现，在人事决策过程中，相比于人类决策，AI 决策会导致个体感知信息去情境化的程度更高，进而降低个体的公平感知。类似地，Nobel 等(2021)发现，应聘者认为 AI 不能读懂“言外之意”(read between the lines)，AI 简历筛选的工作预测相关性(job relatedness-predictive)、工作内容相关性(job relatedness-content)显著低于人类筛选流程，进而对 AI 决策的公平感知更低。Nagtegaal(2021)通过对 AI 和人类决策在不同复杂度的决策情境的公平感知差异的研究，进一步证实了个体对 AI 决策的简化属性的知觉对 AI 和人类决策的公平感知的差异影响作用。具体地，个体认为 AI 不具备从实践中获得隐性知识的能力，决策过程是简化的，因此在简单的决策任务中，个体认为 AI 决策更公平；但是在复杂的决策任务中，个体认为人类决策更公平。

3.3 客观属性 vs. 主观属性

个体对 AI 决策的客观性知觉是影响其对 AI 和人类决策公平感知差异的另一重要内在机制。相比于人类决策，个体认为 AI 决策过程的输入是客观的历史数据和事实，遵循着固定的算法、模型和规则(Lindebaum & Ashraf, in press)，因此，个体往往认为 AI 决策具有更高的一致性、客观性，并且会导致更少的责任归因，进而认为 AI 决策比人类决策更公平。

3.3.1 一致性

AI 决策被认为可以减少人类决策过程的主观性和个体偏见，能够保证决策过程在任意时间点、对所有接受者都是标准化的，进而带来更高的公平感知(Howard et al., 2020)。例如，在招聘决策中，应聘者感知到的 AI 决策的一致性显著高于人类决策，进而认为 AI 决策更公平(Langer, König, & Papathanasiou, 2019; Langer, König, Sanchez, & Samadi, 2019)。

3.3.2 中立性

AI 决策被认为是没有主观意图和个体偏好的，即个体认为 AI 决策是中立的，因此认为其更公平(Miller & Keiser, 2021)。Miller 和 Keiser (2021)比较了交通违法识别中使用 AI 决策和雇佣白人警察决策的公平感知。研究发现，黑人参与者认为，相比于白人警察决策，AI 决策是更加中立和客

观的，进而认为 AI 决策是更公平的。在此基础上，部分研究探讨了决策情境在其中的调节作用(Marcinkowski et al., 2020)。例如，由于高影响力、机械化的决策情境本身对客观性的要求更高，因此，在高影响力的决策情境中，个体会认为 AI 决策更加公平；而在低影响力决策情境中，个体对 AI 和人类决策的公平感知并没有显著差异(Marcinkowski et al., 2020)。

3.3.3 责任归因

当面对决策产生不利的结果时，个体更倾向于将人类决策归咎于决策者的主观意图，从而降低公平感(Brockner et al., 2007)；而当 AI 做决策时，个体认为 AI 是没有主观意图和动机的，无法将其作为道德主体进行评判。因此，个体不会将责任归因于 AI，进而认为 AI 决策更公平。例如，宋晓兵和何夏楠(2020)发现，当消费者面对不公平的价格决策时，他们对人类决策比 AI 决策有更高的蓄意性归因，即他们更可能认为销售人员是故意抬高产品价格而 AI 不是故意的，从而认为 AI 决策比销售人员决策更公平。

3.4 小结

综合来看，现有 AI-人类二元决策的公平感知研究还没有达成一致的结论。部分研究发现，相比于人类决策，个体认为 AI 决策是机械的(缺少情感和互动)和简化的(去情景化)，因此人类决策更加公平(例如, Newman et al., 2020; Nørskov et al., 2020)；另一部分研究则发现，相比于人类决策，个体认为 AI 决策是客观的(一致性、中立性和低责任归因)，因此，AI 决策更加公平(例如, 宋晓兵, 何夏楠, 2020; Langer, König, & Papathanasiou, 2019; Marcinkowski et al., 2020)；同时，还存在少量研究发现，个体对于 AI 和人类决策的公平感知没有显著差异(Ötting & Maier, 2018; Suen et al., 2019)。上述不一致的研究发现主要是由于个体在对决策过程形成知觉时，所聚焦的 AI 属性存在差异(即机械属性、简化属性、客观属性)，而个体究竟基于 AI 的哪种属性来进行判断可能在很大程度上取决于具体的决策情境和个人特征(宋晓兵, 何夏楠, 2020; Kaibel et al., 2019)。可以看出，这一类研究主要是结合传统公平研究相关理论和 AI 相关理论，来探讨 AI 和人类这两类决策主体如何影响个体对决策主体属性知觉的差异，进而影响其对不同决策主体的公平感知。这类研究拓展了现有公平研究中对决策主体类型如何影响公平感知的相关理论。

表 2 AI-人类二元决策的公平感知研究总结

类别	文献数量	机制类别	机制	作者，年份
机械属性 vs. 社会属性	11	情感	情感/人情味/善意	Helberger et al., 2020[†]; Kaibel et al., 2019; Langer et al., in press; Lee & Rich, 2021[†]

		互动	互动性/人际接触/尊重	Acikgoz et al., 2020[†]; Lee & Baykal, 2017[†]; Noble et al., 2021; Nørskov et al., 2020[†]; Ötting & Maier, 2018[†]; Schlicker et al., in press; Wang, 2018
简化属性 vs. 复杂属性	5	去情景化	去情景化/定量化/隐性知识/简化性	Höddinghaus et al., 2021[†]; Lee, 2018[†]; Nagtegaal, 2021[†]; Newman et al., 2020
客观属性 vs. 主观属性	6	一致性	一致性	Howard et al., 2020; Langer, König, & Papathanasiou, 2019; Langer, König, Sanchez, & Samadi, 2019
		中立性	中立性	Marcinkowski et al., 2020[†]; Miller & Keiser, 2021[†]
		责任归因	蓄意性归因	宋晓兵, 何夏楠, 2020

注：[†]指该论文使用了相关逻辑阐述机制，但未进行实证机制测量和检验。

结合以上研究分类和文献梳理，本文总结了 AI 决策公平感知的理论框架，如图 1 所示。

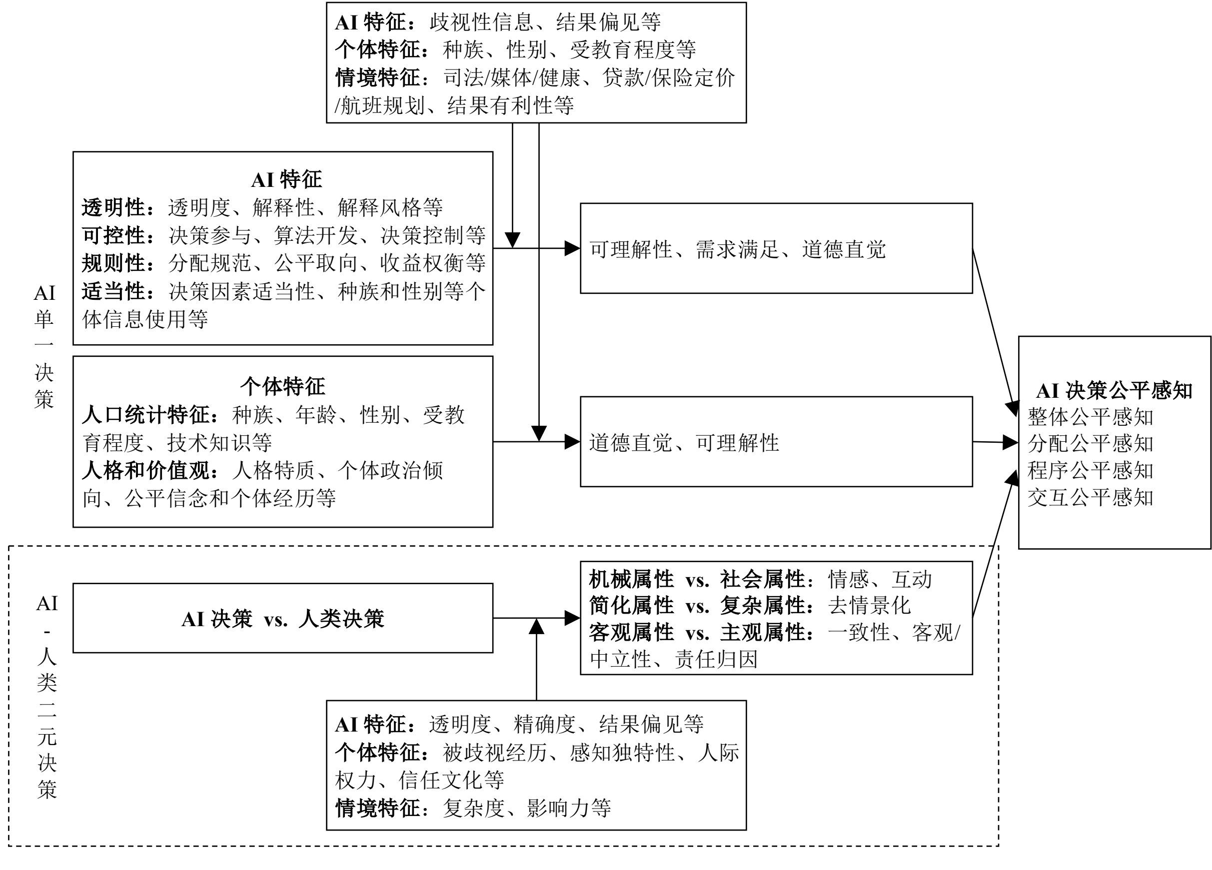


图 1 AI 决策公平感知的影响因素和作用机制

4 理论机制

在前文梳理的基础上，本文进一步提炼了 AI 决策公平感知相关研究的理论机制。值得指出的

是，由于不同的学科对理论的应用程度差异较大，现有研究中明确使用理论建立和阐释假设的相对较少，因此，本文着重梳理了相关研究中相似的逻辑，以及逻辑背后的代表性理论机制。为了更好地展示出各类研究的差异，本文基于前文的研究分类，梳理了以下使用频率较高或具有启发性的理论机制。

4.1 AI 单一决策的公平感知相关理论

4.1.1 道德基础理论

道德基础理论(moral foundations theory; Graham et al., 2013)认为，道德判断是由快速的道德直觉引起的，具有普适性；但同时对于某一事件道德程度的判断又受到个体特征的影响，具有一定的可变性(variability; Graham et al., 2011; Haidt, 2001)。也就是说，不同个体可能支持同一个道德基础(例如，人们普遍认为性别歧视是不道德、不公平的)，但对该事件(不)道德程度的判断又存在个体差异(例如，不同个体对性别歧视不道德程度的判断是有差异的；Graham et al., 2011; Graham et al., 2013)。

基于道德基础理论，一方面，个体根据潜在的、普适的、直观性的公平直觉，会对基于道德相关信息(例如，性别、种族等)的 AI 决策进行公平判断。因此，道德基础理论被用于解释 AI 特征对个体公平感知的影响，并认为其具有普遍性。例如，Nyarko 等(2020)的研究发现，用 AI 对刑事被告做取保候审的风险评估时，个体对使用种族或性别信息的 AI 决策的公平感知较低，因为个体普遍认为基于种族和性别的判断是不公平的(Graham et al., 2011)。

另一方面，个体对基于道德相关信息(例如，性别、种族等)的 AI 决策的公平感知会受到其自身特征的影响。因此，道德基础理论也被用于解释个体特征对 AI 决策公平感知的影响过程，并认为个体的人口统计学特征、人格和价值观等会影响其对 AI 决策的公平感知(Grgić-Hlača et al., 2020)。例如，研究发现，女性比男性更注重伦理道德，因此，女性对包含了可能产生偏见的性别等数据特征的 AI 决策的公平感知更低 (Pierson et al., 2017)。

4.1.2 公平启发理论

公平启发理论(fairness heuristic theory; Lind, 2001)认为，为了减少认知负荷、提高判断效率，个体会利用自己接触到的公平信息，启发式地帮助个体进行公平判断(Qin, Ren, et al., 2018)。尤其是在信息缺乏或模糊的情境下，个体更倾向于利用从情境中获取的不完整的、与公平有关的信息，并结合已有知识信念来迅速地形成一个对所处情境的整体公平感知(李晔 等, 2002)。

基于公平启发理论，AI 作为一种新兴的决策主体，常表现为信息模糊或缺乏，这可能会增加个体的不确定性或不安全感(Acikgoz et al., 2020)，导致个体倾向于采用启发式来做出公平判断，从而降低对 AI 决策的公平感知(Grgić-Hlača, Redmiles, et al., 2018)。因此，公平启发理论被用于指导

和解释信息缺乏或模糊的不确定性情境下个体对 AI 决策的公平感知(Lind, 2001)。例如，Grgić-Hlača 和 Redmiles 等(2018)研究发现，个体会依赖其对 AI 决策的某个特征的内在属性的隐性或显性的评估(即公平启发式)，来形成对 AI 决策的公平判断。

4.1.3 公平理论

公平理论(fairness theory; Adams, 1965)认为，公平感知是基于自我与他人之间的比较、或者过去和现在的自我比较而产生的。这种比较下的不一致性感知会影响个体采用的公平判断规则取向(例如，公平规则、平等规则或需求规则)，进而影响对不同规则取向下决策的公平感知。

根据公平理论，AI 决策中的公平性规则与个体自身的公平规则取向会形成比较，进而影响个体对 AI 决策的公平感知(Shin, 2010; Shin & Park, 2019)。尤其是当 AI 决策的规则与个体公平取向一致时，个体的公平感知更高。因此，公平理论主要用于解释，当个体对于 AI 决策的公平性规则认知与自身公平取向(不)一致性时，对 AI 单一决策公平感知的影响。例如，Lee 等(2019)发现，当个体公平取向与 AI 决策的公平取向不同时，个体认识到 AI 决策的固有局限性，从而降低了公平感知。

4.1.4 小结

综上所述，AI 单一决策的公平感知研究的相关理论，主要基于传统公平感知理论的三种视角，即道德义务视角、信息加工视角和社会交换视角(李晔 等, 2002; Colquitt & Zipay, 2015)，来解释 AI 特征和个体特征对 AI 决策公平感知的影响。具体来说，道德基础理论基于道德义务视角，认为个体通过普遍认可的道德规范来判断 AI 决策的公平性，这类研究主要侧重于解释，AI 适当性和个体特征通过道德直觉机制影响 AI 公平感知的过程。公平启发理论基于信息加工视角，认为个体面对有限的信息时会通过启发式思维来判断 AI 决策的公平性，这类研究主要侧重于解释，AI 透明性通过可理解性机制影响 AI 公平感知的过程。公平理论则基于社会交换视角，认为个体通过 AI 决策中的经济和社会交换是否符合自己的需要来判断 AI 决策的公平性，这类研究主要侧重于解释，AI 可控性和规则性通过需求满足机制影响 AI 决策公平感知的过程。

4.2 AI-人类二元决策的公平感知研究的相关理论

4.2.1 计算机作为社会参与者理论

计算机作为社会参与者理论(computers are social actors CASA; Nass & Moon, 2000)认为，人们对计算机的反应是“社会性的”，即人们在与计算机交互时，会将人与人交互过程中的社会规则、规范和期望运用到人-机交互情境中，并相应地表现出和人-人交互中相同的社交反应。因此，这个理论被广泛运用在人-机交互的研究中，用来解释和预测人们对技术(例如，AI 等)的情感和行为反应(例如，Qin et al., 2021; Qin et al., 2022)。

AI-人类二元决策的公平感知研究，事实上是构建了人-人交互情境与人-AI 交互情境中的个体公平感知的对比。因此，CASA 理论主要应用于解释 AI 作为决策主体时个体的反应，并认为个体在上述两种交互情境中的心理和行为反应是相似的。例如，基于 CASA 理论，Suen 等(2019)探究了人们对 AI 决策与人类决策的整体公平感知的异同，并发现人们对上述两种决策的公平感知是相似的。

4.2.2 简化理论

简化理论(theory of reductionism; Choi et al., 2007; Nisbett et al., 2001; Von Bertalanffy, 1972)认为，当外部世界的信息被量化(quantifying)处理时，非量化的质性信息会被删除或被简单表征。这种被简化的、去情境化的(decontextualization)感知会引发个体的消极反应。

根据简化理论，相比于人类决策，AI 决策的量化特征会让个体感知决策过程是简化和去情境化的。因此，简化理论主要用于解释个体对 AI 决策信息处理过程的知觉与感受，并认为个体知觉到 AI 与人类二者决策的信息处理过程存在差异，进而影响其公平感知。例如，基于简化理论，Newman 等(2020)发现，在人事决策过程中，相比于人类决策，AI 决策会导致应聘者感知去情境化的程度更高，进而降低应聘者对 AI 决策的公平感知。

4.2.3 机器启发理论

机器启发理论(machine heuristic model; Sundar, 2008)认为，机器比人类更安全、更值得信赖，即当个体认为与其交互的是一台机器而不是人类时，个体会自动地启动关于机器的刻板印象，即个体会认为它是客观的、意识形态上无偏见的等，从而引发个体的反应。

根据机器启发理论，个体会认为 AI 决策比人类决策更客观与公平。因此，机器启发理论主要被用于解释个体对 AI 自动化的刻板印象和反应，并认为个体对 AI 与人类两者的刻板印象与反应存在差异(Araujo et al., 2020)。例如，基于机器启发理论，Helberger 等(2020)发现，当交互界面提示用户正在与 AI 而不是人类打交道时，就会触发用户的机器启发式思维，使得用户认为 AI 的决定和选择是客观的、没有偏见的，进而认为 AI 决策比人类决策更为公平。

4.2.4 小结

综上所述，AI-人类二元决策公平感知的研究，主要是基于个体知觉的 AI 和人类这两类决策主体的属性差异，来解释个体对不同主体所做出的决策的公平感知差异。具体来说，CASA 侧重于 AI 的社会属性，认为个体在人-人交互情境与人-AI 交互情境中的公平感知是相似的。简化理论侧重于 AI 的简化属性，认为 AI 决策的量化特征会让个体感知人类决策比 AI 决策更公平。机器启发理论则侧重于 AI 的客观属性，认为 AI 是客观的、无偏见的刻板印象会让个体感知 AI 决策比人类决策更公平。可以看出，AI-人类二元决策公平感知研究尚未达成一致结论，这可能是因为个人在不同的情境下所关注的、用于与人类决策对比的 AI 属性不同。

5 研究总结与未来研究方向

通过对现有文献的全面回顾和对理论机制的梳理，可以发现，尽管 AI 决策公平感知的相关研究已经引起了多学科领域的广泛关注，但仍然较为松散，缺乏系统性，也存在以下不足。具体来说，在 AI 单一决策的公平感知研究方面，第一，现有研究主要关注个体如何通过认知过程(例如，可理解性等)形成对 AI 决策的公平感知，缺乏对相关情绪机制的关注；第二，现有研究更多聚焦于个体对 AI 决策在分配和程序公平方面的感知，缺乏对交互公平感知的关注；第三，现有研究主要关注个体对算法类 AI 的决策公平感知，缺乏对实体类 AI(例如，机器人型 AI)决策公平感知的探索。在 AI-人类二元决策的公平感知研究方面，第四，现有研究尚未形成统一结论，缺乏对边界条件的探索；第五，现有研究仅比较了 AI 和人类作为决策主体时的公平感知差异，缺乏对 AI 和人类在决策中更为复杂的情况(例如，共同决策等)的探讨。基于上述不足，本文总结了以下未来研究方向。

5.1 探索 AI 决策公平感知的情绪影响机制

现有 AI 单一决策公平感知的相关研究主要关注个体如何通过认知过程形成对 AI 决策的公平感知，缺乏对情绪机制的探索。组织公平相关研究指出, 个体的公平感知及其影响因素之间存在认知和情绪的双路径作用机制(Colquitt & Zipay, 2015)。缺乏对情绪影响机制的研究，不利于对 AI 决策公平感知形成系统和全面的认识。因此，未来研究有必要系统地探讨 AI 决策公平感知的情绪影响机制。例如，现有研究发现，外界事物相关信息的模糊性和不确定性会通过引发个体的消极情绪(例如，惊恐)，进而降低其公平感知(陈晨, 秦昕 等, 2020; 王芹 等, 2012; Tene & Polonetsky, 2015)。AI 作为一种新兴的决策主体，常表现为信息模糊或缺乏(Langer et al., 2018)，这可能会诱发个体的消极情绪，降低其对 AI 决策的公平感知。综上，未来研究可以从情绪的视角来进一步揭示 AI 决策公平感知的影响机制。

5.2 探索 AI 决策交互公平感知的影响因素

现有关于 AI 决策公平感知的相关研究主要关注三种公平类型中的程序公平和结果公平，对交互公平的关注较为缺乏。因此，未来研究可以考察 AI 决策交互公平感知的影响因素，以形成全面系统地了解。例如，相比于人类，AI 因为缺乏人际互动性等(Glikson & Woolley, 2020)，可能会导致个体感知到比较低的交互公平。进一步，AI 决策交互公平感知的影响因素可能会存在一些边界条件。例如，AI 类型可能是一个重要的边界条件，当决策主体为嵌套式 AI 和虚拟式 AI 时，个体对 AI 决策的交互公平感知要低于人类决策；但是当决策主体为机器人形 AI 时，个体对 AI 决策的交互公平感知可能与人类决策的差异较小。

5.3 探索机器人形 AI 决策公平感知的影响因素

现有研究主要关注嵌套式 AI 和虚拟式 AI 决策的公平感知(例如，Acikgoz et al., 2020; Newman et al., 2020)，缺乏对机器人形 AI 决策公平感知的关注。机器人形 AI 具有高拟人化(anthropomorphism)、高人际亲密性(immediacy)、可触摸性(tangibility)等特点(Glikson & Woolley, 2020)，这些因素都可能会影响个体对 AI 决策的公平感知过程(Acikgoz et al., 2020)。基于此，未来研究可以进一步探讨个体对机器人形 AI 决策的公平感知。例如，已有研究发现，嵌套式 AI 和虚拟式 AI 具有较低的人际互动性、开放性和双向沟通(Acikgoz et al., 2020)，导致个体对 AI 决策的公平感知较低。但机器人形 AI 在决策过程中具有较高的人际互动性、开放性和双向沟通，可能会增强个体的情感信任，进而产生较高的公平感知。进一步，未来研究还可以探索个体对不同形式 AI 决策的公平感知差异。综上，探索个体对机器人形 AI 决策的公平感知不仅能为现有 AI 决策公平感知的相关研究提供更全面的认识，还能为管理实践提供更全面和有效的指引。

5.4 探索影响 AI-人类二元决策的公平感知的边界条件

现有 AI-人类二元决策的公平感知研究尚未形成一致结论，即部分研究发现 AI 决策更公平(例如，宋晓兵, 何夏楠, 2020)，另一部分研究则发现人类决策更公平(例如，Newman et al., 2020)，这表明可能存在非常重要的边界条件。例如，决策内容的可量化程度可能是一个重要的边界条件。当决策内容可量化程度较低时，人类决策的准确性更高，个体可能对人类决策的公平感知更高；相反，当决策内容可量化程度较高时，AI 决策的准确性更高，个体可能对 AI 决策的公平感知更高。其次，中西方文化差异也可能是一个潜在的边界条件。由于文化和价值观的不同，西方文化情境中得到的一些研究结论可能无法推广到中国本土情境中。综上，通过对边界条件的探索，有助于进一步厘清现有研究结果出现不一致的原因，并对现有研究结论进行有效整合。

5.5 探索复杂情景中 AI 和人类共同决策对公平感知的影响

在探索 AI-人类二元决策的公平感知研究中，目前研究仅仅比较了 AI 和人类作为决策主体时的差异，缺乏对 AI 和人类在决策中更为复杂情形的探讨。一个决策可能是 AI 和人类共同完成的，在这个过程中，AI 和人类所承担的不同角色(例如，在辅助决策中，人类辅助 AI 决策、AI 辅助人类决策等)、二者的决策顺序(例如，AI 先决策人类后决策、人类先决策 AI 后决策等)等都可能对个体的公平感知产生影响。事实上，AI 和人类共同决策的现象已经越来越多地涌现出来(Newman et al., 2020)。例如，肿瘤科医生在对病人病情进行诊断时，AI 综合各方面的检查结果来辅助医生进行诊断决策；在人员招聘的多轮选拔过程中，AI 负责第一轮海选筛查简历，推荐出进入第二轮由人类来面试的候选应聘者，AI 和人类决策共同决定着最终的招聘结果。而这些不同的决策情形如何影响个体的公平感知，目前还未有研究进行回答。以辅助决策为例，相比于人类辅助 AI 决策的情形，AI

辅助人类决策可能会让个体感知到更高的公平感，因为人类可以在 AI 提供的客观、量化信息基础上进行综合分析与决策(Sundar, 2008)。综上，通过对更为复杂决策情形的探索，有助于全面地理解个体对 AI 与人类共同决策过程中的公平感知。

参考文献

曹培杰. (2020). 人工智能教育变革的三重境界. *教育研究, 481*, 143–150.

陈晨，秦昕，谭玲，卢海陵，周汉森，宋博迪. (2020). 授权型领导—下属自我领导匹配对下属情绪衰竭和工作绩效的影响. *管理世界, 36*(12), 145–162.

陈晨，张昕，孙利平，秦昕，邓惠如. (2020). 信任以稀为贵？下属感知被信任如何以及何时导致反生产行为. *心理学报, 52*(3), 329–344.

房鑫，刘欣. (2019). 论人工智能时代人力资源管理面临的机遇和挑战. *山东行政学院学报, 167*, 104–109.

郭秀艳，郑丽，程雪梅，刘映杰，李林. (2017). 不公平感及相关决策的认知神经机制. *心理科学进展, 25*(6), 903–911.

李超平，时勘. (2003). 分配公平与程序公平对工作倦怠的影响. *心理学报, 35*(5), 677–684.

李晔，龙立荣，刘亚. (2002). 组织公平感的形成机制研究进展. *人类工效学, 8*(1), 38–41.

秦昕，薛伟，陈晨，刘四维，邓惠如. (2019). 为什么领导做出公平行为：综述与未来研究方向. *管理学季刊, 4*(4), 39–62.

宋晓兵，何夏楠. (2020). 人工智能定价对消费者价格公平感知的影响. *管理科学, 33*(5), 3–16.

王芹，白学军，郭龙健，沈德立. (2012). 负性情绪抑制对社会决策行为的影响. *心理学报, 44*(5), 690–697.

吴燕，周晓林. (2012). 公平加工的情境依赖性：来自 ERP 的证据. *心理学报, 44*(6), 797–806.

谢洪明，陈亮，杨英楠. (2019). 如何认识人工智能的伦理冲突?——研究回顾与展望. *外国经济与管理, 41*(10), 109–124.

谢小云，左玉涵，胡琼晶. (2021). 数字化时代的人力资源管理：基于人与技术交互的视角. *管理世界, 37*(1), 200–216+13.

徐鹏，徐向艺. (2020). 人工智能时代企业管理变革的逻辑与分析框架. *管理世界, 36*(1), 122–129.

杨文琪，金盛华，何苏日那，张潇雪，范谦. (2015). 非人化研究：理论比较及其应用. *心理科学进展, 23*(7), 1267–1279.

张志学，赵曙明，施俊琦，秦昕，贺伟，赵新元,… 吴刚. (2021). 数字经济下组织管理研究的关键科学问题——第 254 期“双清论坛”学术综述. *中国科学基金, 35*(5), 774-781.

郑功成. (2009). 中国社会公平状况分析——价值判断、权益失衡与制度保障. *中国人民大学学报, 23*(2), 2–11.

周浩，龙立荣. (2007). 公平敏感性研究述评. *心理科学进展, 15*(4), 702–707.

Acikgoz, Y., Davison, K. H., Compagnone, M., & Laske, M. (2020). Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment, 28*(4), 399–416.

Adams, J. S. (1965). Inequity in social exchange. *Advances in Experimental Social Psychology, 2*, 267–299.

Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society, 35*(3), 611–623.

Balasubramanian, N., Ye Y., & Xu M. (in press). Substituting human decision-making with machine learning: Implications for organizational learning. *Academy of Management Review.* Advance online publication. https://doi.org/10.5465/amr.2019.0470

Barabas, C., Doyle, C., Rubinovitz, J., & Dinakar, K. (2020, January). *Studying up: Reorienting the study of algorithmic fairness around issues of power.* Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34.

Bigman, Y. E., Yam, K. C., Marciano, D., Reynolds, S. J., & Gray, K. (in press). Threat of racial and economic inequality increases preference for algorithm decision-making. *Computers in Human Behavior.* Advance online publication. https://dx.doi.org/10.1016/j.chb.2021.106859

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018 April). *'It's reducing a human being to a percentage'; Perceptions of justice in algorithmic decisions.* Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montréal, Canada.

Brockner, J., Fishman, A. Y., Reb, J., Goldman, B., Spiegel, S., & Garden, C. (2007). Procedural fairness, outcome favorability, and judgments of an authority's responsibility. *Journal of Applied Psychology, 92*(6), 1657–1671.

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*(2), 220–239.

Chang, M. L., Pope, Z., Short, E. S., & Thomaz, A. L. (2020 August). *Defining fairness in human-robot teams.* Proceedings of 2020 29th IEEE International Conference on Robot and Human Interactive Communication, Virtual Conference.

Cheng, H. F., Stapleton, L., Wang, R., Bullock, P., Chouldechova, A., Wu, Z. S. S., & Zhu, H. (2021, May). *Soliciting stakeholders' fairness notions in child maltreatment predictive systems*. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan.

Choi, I., Koo, M., & Choi, J. A. (2007). Individual differences in analytic versus holistic thinking. *Personality and Social Psychology Bulletin, 33*(5), 691–705.

Colquitt, J. A., & Zipay, K. P. (2015). Justice, fairness, employee reactions. *Annual Review of Organizational Psychology and Organizational Behavior, 2*(1), 75–99.

Dalenberg, D. J. (2018). Preventing discrimination in the automated targeting of job advertisements. *Computer Law & Security Review, 34*(3), 615–627.

Deutsch, M. 1975. Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues, 31*(3), 137–149.

Dodge, J., Vera Liao, Q., Bellamy, & R. K. E. (2019 March). *Explaining models: An empirical study of how explanations impact fairness judgment.* Proceedings of the International Conference on Intelligent User Interfaces, Marina del Rey, CA.

Fischhoff, B., & Broomell, S. B. (2020). Judgment and Decision Making. *Annual Review of Psychology, 71*, 331–355.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine, & A. Plant (Eds.), *Advances in experimental social psychology* (Vol. 47, pp. 55–130). New York: Academic Press.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101*, 366–385.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*, 619–619.

Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018 April). *Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction.* Proceedings of the 2018 World Wide Web Conference on World Wide Web, Lyon, France.

Grgić-Hlača, N., Weller, A., & Redmiles, E. M. (2020 November). *Dimensions of diversity in human perceptions of algorithmic fairness*. Proceedings of the CSCW 2019 Workshop on Team and Group Diversity, Austin, Texas.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018 February). *Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning*. Proceedings of the 32th AAAI Conference on Artificial Intelligence, New Orleans, Louisiana.

Haidt, J. (2001). The emotional dog and its rationalist tail: A social intuitionist approach to moral judgment. *Psychological*

Review, 108(4), 814–834.

Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., & Ur, B. (2020 January). *An empirical study on the perceived fairness of realistic, imperfect machine learning models.* Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency, Barcelona, Spain.

Helberger, N., Araujo, T., & de Vreese, C. H. (2020). Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review, 39*, Article 105456. https://doi.org/10.1016/j.clsr.2020.105456

Höddinghaus, M., Sondern, D., & Hertel, G. (2021). The automation of leadership functions: Would people trust decision algorithms?. *Computers in Human Behavior, 116*, Article 106635. https://doi:10.1016/j.chb.2020.106635

Howard, F. M., Gao, C. A., & Sankey, C. (2020). Implementation of an automated scheduling tool improves schedule quality and resident satisfaction. *Plos One, 15*(8), Article e0236952. https://doi.org/10.1371/journal.pone.0236952

Htun, N. N., Lecluse, E., & Verbert, K. (2021, April). *Perception of fairness in group music recommender systems.* In 26th International Conference on Intelligent User Interfaces, College Station, TX, USA.

Hutchinson, B., & Mitchell, M. (2019, January). *50 years of test (un) fairness: Lessons for machine learning.* Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA.

Kaibel, C., Koch-Bayram, I., Biemann, T., & Mühlenbock, M. (2019 July). *Applicant perceptions of hiring algorithms-uniqueness and discrimination experiences as moderators.* Proceedings of the Academy of Management Annual Meeting, Briarcliff Manor, NY.

Karam, E. P., Hu, J., Davison, R. B., Juravich, M., Nahrgang, J. D., Humphrey, S. E., & Scott DeRue, D. (2019). Illuminating the 'face' of justice: A meta-analytic examination of leadership and organizational justice. *Journal of Management Studies, 56*(1), 134–171.

Kasinidou, M., Kleanthous, S., Barlas, P., & Otterbacher, J. (2021, March). *I agree with the decision, but they didn't deserve this: Future developers' perception of fairness in algorithmic decisions*. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada.

Langer, M., König, C. J., & Fitili, A. (2018). Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection. *Computers in Human Behavior, 81*, 19–30.

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment, 27*(3), 217–234.

Langer, M., König, C. J., Back, C., & Hemsing, V. (in press). Trust in Artificial Intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. PsyArXiv Prepints. https://doi.org/10.31234/osf.io/r9y3t

Langer, M., König, C. J., Sanchez, D. R. P., & Samadi, S. (2019). Highly automated interviews: Applicant reactions and the organizational context. *Journal of Managerial Psychology, 35*(4), 301–314.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1), Article 2053951718756684. https://doi:10.1177/2053951718756684

Lee, M. K., & Baykal, S. (2017 February). *Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division.* Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. Portland, OR.

Lee, M. K., & Rich, K. (2021 May). *Who is included in human perceptions of AI? Trust and perceived fairness around healthcare AI and cultural mistrust.* Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Yokohama, Japan.

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer*

Interaction, 3, 182–208.

Leventhal, G. S. (1976). The distribution of rewards and resources in groups and organizations. In L. Berkowitz, & E. Walster (Eds.), *Advances in experimental social psychology* (Vol. 9, pp. 91–131). New York: Academic Press.

Lind, E. A. (2001). Fairness heuristic theory: Justice judgments as pivotal cognitions in organizational relations. In J. Greenberg & R. Cropanzano (Eds.), *Advances in organizational justice*, (Vol. 1, pp. 56–88). Stanford, CA: Stanford university press.

Lindebaum, D., & Ashraf, M. (in press). The ghost in the machine, or the ghost in organizational theory? A complementary view on the use of machine learning. *Academy of Management Review.* https://doi.org/10.5465/amr.2021.0036

Lindebaum, D., Vesa, M., & Den Hond, F. (2020). Insights from "The Machine Stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Academy of Management Review, 45*(1), 247–263.

Loehr, A. Big data for HR: Can predictive analytics help decrease discrimination in the workplace? *The Huffington Post*. Retrieved March 23, 2015, from https://www. huffingtonpost.com

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650.

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020 January). *Implications of AI (un-) fairness in higher education admissions: The effects of perceived AI (un-) fairness on exit, voice and organizational reputation.* the Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain.

Martínez-Miranda, J., & Aldea, A. (2005). Emotions in human and artificial intelligence. *Computers in Human Behavior, 21*(2), 323–341.

Miller, S. M., & Keiser, L. R. (2021). Representative bureaucracy and attitudes toward automated decision making. *Journal of Public Administration Research and Theory, 31*(1), 150–165.

Nagtegaal, R. (2021). The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly, 38*(1). Article 101536. https://doi:10.1016/j.giq.2020.101536

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103.

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes, 160*, 149–167.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*(2), 291–310.

Noble, S. M., Foster, L. L., & Craig, S. B. (2021). The procedural and interpersonal justice of automated application and resume screening. *International Journal of Selection and Assessment*, Advance online publication. https://doi.org/10.1111/ijsa.12320

Nørskov, S., Damholdt, M. F., Ulhøi, J. P., Jensen, M. B., Ess, C., & Seibt, J. (2020). Applicant fairness perceptions of a robot–mediated job interview: A video vignette–based experimental survey. *Frontiers in Robotics and AI, 7*, Article 586263. https://doi.org/10.3389/frobt.2020.586263

Nyarko, J., Goel, S., & Sommers, R. (2020 October). *Breaking taboos in fair machine learning: An experimental study*. (Unpublished doctorial dissertation). Stanford University.

Ötting, S. K., & Maier, G. W. (2018). The importance of procedural justice in human-machine interactions: Intelligent systems as new decision agents in organizations. *Computers in Human Behavior, 89*, 27–39.

Pierson, E. (2017). *Gender differences in beliefs about algorithmic fairness.* arXiv preprint. http://arxiv.org/abs/1712.09124v2

Pierson, E. (2018). *Demographics and discussion influence views on algorithmic fairness.* arXiv preprint.

http://arxiv.org/abs/1712.09124

Plane, A. C., Redmiles, E. M., Mazurek, M. L., Tschantz, M. C., & Assoc, U. (2018 August). *Exploring user perceptions of discrimination in online targeted advertising.* Proceedings of the 26th USENIX Security Symposium, Vancouver, BC.

Qin, X., Chen, C., Yam, K. C., Cao, L., Li, W., Guan, J., Zhao, P., Dong, X., & Lin, Y. (2022). Adults still can't resist: A social robot can induce normative conformity. *Computers in Human Behavior, 127*. Article 107041. https://doi.org/10.1016/j.chb.2021.107041

Qin, X., Huang, M., Johnson, R. E., Hu, Q., & Ju, D. (2018). The short-lived benefits of abusive supervisory behavior for actors: An investigation of recovery and work engagement. *Academy of Management Journal, 61*(5), 1951–1975.

Qin, X., Ren, R., Zhang, Z., & Johnson, R. E. (2015). Fairness heuristics and substitutability effects: Inferring the fairness of outcomes, procedures, and interpersonal treatment when employees lack clear information. *Journal of Applied Psychology, 100*(3), 749–766.

Qin, X., Ren, R., Zhang, Z., & Johnson, R. E. (2018). Considering self-interests and symbolism together: How instrumental and value-expressive motives interact to influence supervisors' justice behavior. *Personnel Psychology, 71*(2), 225–253.

Qin, X., Yam, K. C., Chen, C., & Li, W. (2021). Revisiting social robots and their impacts on conformity: Practical and ethical considerations. *Science Robotics, eLetters*. Retrieved October 25, 2021, from https://www.science.org/doi/full/10.1126/scirobotics.aat7111.

Rupp, D. E., & Cropanzano, R. (2002). The mediating effects of social exchange relationships in predicting workplace outcomes from multifoci organizational justice. *Organizational behavior and human decision processes*, *89*(1), 925–946.

Saha, D., Schumann, C., Mcelfresh, D., Dickerson, J., Mazurek, M., & Tschantz, M. (2020, November). *Measuring non-expert comprehension of machine learning fairness metrics.* Proceedings of International Conference on Machine Learning, Online Conference.

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2020). How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. *Artificial Intelligence, 283*, Article 103238. https://doi.org/10.1145/3306618.3314248

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70.

Schlicker, N., Langer, M., Ötting, S., Baum, K., König, C. J., & Wallach, D. (in press). What to expect from opening up 'Black Boxes'? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior.* Advance online publication. https://doi.org/10.1016/j.chb.2021.106837

Schoeffer, J., Machowski, Y., Kuehl N. (2021 April). *A study on fairness and trust perceptions in automated decision making.* Proceedings of the ACM IUI 2021 Workshops, College Station, USA.

Shin, D. (2010). The effects of trust, security and privacy in social networking: A security-based approach to understand the pattern of adoption. *Interacting with Computers, 22*(5), 428–438.

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media, 64*(4), 541–565.

Shin, D. (2021a). A cross-national study on the perception of algorithm news in the East and the West. *Journal of Global Information Management, 29*(2), 77–101.

Shin, D. (2021b). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, Article 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shin, D. (in press). The perception of humanness in conversational journalism: An algorithmic information-processing perspective. *New Media & Society.* Advance online publication. https://doi.org/10.1177/1461444821993801

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in*

Human Behavior, 98, 277–284.

Smith, Y. N. (2020). *The African American perception of body-worn cameras on police performance and fairness* (Unpublished doctoral dissertation), Capella University, Minneapolis.

Srivastava, M., Heidari, H., & Krause, A. (2019 August). *Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning.* Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage, AK, USA

Suen, H. Y., Chen, Y. C., & Lu, S. H. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior, 98*, 93–101.

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger, & Flanagin, A. J. (Eds.) *Digital media, youth, and credibility* (pp. 73–100). Cambridge, MA: MIT Press.

Tene, O., & Polonetsky, J. (2015). A theory of creepy: Technology, privacy, and shifting social norms. *Yale Journal of Law and Technology, 16*(1), 59–102.

Uhde, A., Schlicker, N., Wallach, D. P., & Hassenzahl M. (2020 April). *Fairness and decision-making in collaborative shift scheduling Systems.* Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI.

van Berkel, N., Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R. M., & Kostakos, V. (2019). Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction, 3*, 28–46.

van Berkel, N., Goncalves, J., Russo, D., Hosio, S., Skov, M. B. (2021 May). *Effect of information presentation on fairness perceptions of machine learning predictors.* Proceedings in CHI Conference on Human Factors in Computing Systems, Yokohama, Japan.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., … & Nerini, F. F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications, 11*(1), 1–10.

Von Bertalanffy, L. (1972). The history and status of general systems theory. *Academy of Management Journal, 15*, 407–426.

Wang, A. J. (2018). Procedural justice and risk-assessment algorithms. *SSRN Electronic Journal*. Article 3170136. http://dx.doi.org/10.2139/ssrn.3170136

Wang, R., Harper, F. M., & Zhu, H. (2020 April). *Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences.* Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI.

World Social Report. (2020). *Inequality in a rapidly changing world*. Retrieved March 23, 2020, from https://www.un.org/pt/node/89779

The antecedents and underlying mechanisms of fairness perceptions of artificial intelligence decision-making

JIANG Luyuan[1], CAO Limei[1], QIN Xin[1], TAN Ling[2], CHEN Chen[1], PENG Xiaofeng[1]

([1] School of Business, Sun Yat-sen University, Guangzhou 510275, China)

([2] School of Management, Guangdong University of Technology, Guangzhou 510520, China)

Abstract: Inequality is the biggest challenge for global social and economic development, which has the potential to impede the goal of global sustainable development. One way to reduce such inequality is to use artificial intelligence (AI) for decision-making. However, recent research has found that while AI is more accurate and is not influenced by personal bias, people are generally averse to AI decision-making and perceive it as being less fair. Given the theoretical and practical importance of fairness perceptions of AI decision-making, a growing number of researchers have recently begun investigating how individuals form fairness perceptions in regards to AI decision-making. However, existing research is generally quite scattered and disorganized, which has limited researchers' and practitioners' understanding of fairness perceptions of AI decision-making from a conceptual and systematic perspective. Thus, this review first divided the relevant research into two categories. That is, (a) fairness perception studies in which AI is the decision-maker, which focus on how AI characteristics and individual characteristics affect individuals' fairness perceptions; and (b) fairness perception studies that compare AI and humans as decision-makers, which focus on the comparative effects of AI or humans as decision makers on individuals' fairness perceptions. Based on this systematic review, we proposed five promising directions for future research, such as exploring the affective mechanisms underlying the relationship between AI or individual characteristics and fairness perceptions.

Keywords: artificial intelligence, algorithm, fairness, decision-making, antecedents